Master's Thesis

Prediction of unloading crude oil using machine learning with meteorological data.

HYUNSEONG JI

Graduate School of Technology and Innovation Management
UNIST

2020

Prediction of unloading crude oil using machine learning with meteorological data.

HYUNSEONG JI

Graduate School of Technology and Innovation Management
UNIST

Prediction of unloading crude oil using machine learning with meteorological data.

A thesis/dissertation
submitted to the Graduate School of Technology and Innovation
Management, UNIST
in partial fulfillment of the
requirements for the degree of
Master in Technology and Innovation Management

HYUNSEONG JI

12/11/2019

Approved by

Advisor
Han-Gyun Woo

Prediction of unloading crude oil using machine learning with meteorological data.

HYUNSEONG JI

This certifies that the thesis of HYUNSEONG JI is approved.

12/11/2019

Advisor: Han-Gyun Woo

Yeolib Kim

Young-Choon Kim

Abstract

Many machine learning applications are being employed to forecast weather conditions. In this paper, we focus more on small-scale weather forecasts with limited meteorological observation data. When oil refinery companies in non-oil-producing countries import crude oil by VLCCs (Very Large Crude Carriers), VLCCs unload crude oil to onshore storage tanks using SPM (Single Point Buoy Mooring System). Weather conditions in the offshore area where loading buoys are anchored are critical in determining whether unloading process is possible. The current practice of such decision making relies mostly on human experiences, and the predictive accuracy of the current practice is reported as about 75%. We tested machine learning methods to see if these methods can increase predictive accuracy in this problem of classification, the possibility of unloading given weather conditions such as wave heights, wind speeds, and wind directions. The results of our analysis showed that random forest and XGBoost have much better performance (more than 90%) than support vector machines and logistic regression in predicting unloading conditions in the time range from one hour to three days.

Contents

Ⅰ. Introduction ----------1

Ⅱ. Related Work----------3

Ⅲ. Data and Methodology----------6

3.1 Data----------6

3.1.1 Independent Variables----------7
3.1.2 Variables correlation----------9
3.1.3 Dependent Variables----------10

3.2 Methodology ----------10

3.2.1 Logistic Regression----------10
3.2.2 Random Forest ----------11
3.2.3 XGBoost ----------11
3.2.4 K-Fold cross validation ----------12
3.2.5 ROC analysis ----------13
3.2.6 Support Vector Machines ----------14

Ⅳ. Result ----------15

Ⅴ. Conclusion ----------35

1. Introduction

Many downstream oil companies in non-oil-producing countries have imported crude oil by VLCC (Very Large Crude Carriers). However, the voyage from the loading site to the unloading site takes few days to weeks. Meanwhile, it takes about 20 days for oil refinery companies in the Republic of Korea to import crude oil from Saudi Arabia. Thus, the inventory management of crude oil storage tank is fundamental as immediate crude oil supplies are difficult.

Crude oil is unloaded in tanks through buoys after arriving from where it is imported. The unloading takes hours and the crude oil may contain some sea water and sludge. Hence, it takes some days to remove the impurities and drain sea water in order to feed the crude oils to CDUs (Crude Distillation Unit). This indicates that a minimum inventory of crude oil should be secured to feed the CDUs.

The buoys where VLCC unloads the crude oil are affected by sea weather such as wave height, wave direction, wind speed, etc. Moreover, the tension between the floating hose attached to the buoys is high due to bad weather, thereby leading to oil spill accidents. The current practice of such decision making if the VLCC can unload crude oil relies mostly on human experiences rather than by statistical method or data analysis. For example, ship carpenters have realized that crude unloading from VLCC could be difficult under empirically specific conditions above certain levels of wave height (3m). Thus, it has been the current practice of decision making if the VLCC can unload crude oil to storage tank through buoys. This practice is used for crude oil inventory management. The employees responsible for managing crude inventory consider the weather medium-term forecast data in the websites of meteorological agencies for 10 days and forecast if the VLCC could unload crude oil based on the current practice of decision making. The problem is that the weather forecast has high volatility which may lead to a wrong forecast and lower profitability. If the crude oil inventory forecast is lower than the minimum operable inventory in the near future, some amount of reserved crude oil is leased to KNOC (Korea National Oil Company) which incurs costs as rental fee. However, a worst case is when the sea weather suddenly becomes worse unlike medium-term forecast such that the VLCC could not unload the crude oil. It is impossible to solve this problem by leasing reserved crude oil from KNOC due to contracts made with companies a few days before. Thus, the throughput and products output should be reduced. It could be tremendous profit loss.

In this paper, a standard that judges if the VLCC Could unload crude oil or not is set. A forecast model is then established if the VLCC could unload crude oil based on machine learning algorithms such as logistic regression, random forest, XGBoost, and SVM. The meteorological data are imported from https://data.kma.go.kr and the VLCC unloading history data are imported from

https://new.portmis.go.kr and referenced by oil refinery company in the Republic of Korea from January 1, 2016 to December 31, 2018. The buoy used is Ulsan buoy.

2. Related Work

Several researches have forecasted weather using machine learning. However, the weather forecast is highly uncertain. Therefore, for many applications, forecasts are considered valuable only if an uncertainty estimate can be assigned to them [1]. Ingsrisawang and Lily presented some machine learning approaches such as Decision Tree, Artificial Neural Network (ANN) and Support Vector Machines (SVM) for short-term rain forecasting systems [2]. Meteorological datasets are collected during 2004-2006 from Chalermprakiat Royal Rain Making Research Center and classify the rain amount into three classes as no-rain (0-0.1 mm), few-rain (0.1-10 mm) and moderate-rain (>10 mm.) They analyzed other weather conditions and precipitation data and the results are as below. The accuracy of the five-fold cross validation of the decision Tree algorithm is 94.41% while the overall accuracy is 62.57%. However, the accuracy of ANN and SVM are 68.15% and 69.10%, respectively. Meanwhile, the accuracy of SVM is higher than that of ANN and decision tree algorithms. Furthermore, Joshi et al., utilized a Hidden Markov Model (HMM) for the prediction of quantitative snowfall [3]. Model predicts snowfall at different stations in Great Himalayan mountain ranges for two days in advance. The HMM analysis proceeded an analysis that predicted snowfall after two days with nine meteorological variables from 1992 to 2012. The Forward and Viterbi algorithms were used to set the probable observation and state sequence while the Baum-Welch algorithm was used to set the parameter. The prediction accuracy of the HMM on the first day ranged from 60.3% to 79.6% while its prediction accuracy on the second day ranged from 15.9% to 38.2%.

The environmentally-friendly energy industry is also affected by weather. Although the demand for environmentally-friendly energy has increased, the prediction of energy generation is difficult because it is affected by weather. Thus, machine learning is useful for increasing the accuracy of solar generation prediction from weather forecast [4]. The input data are historical weather data and solar intensity observations which enable the computation of future solar intensity. They collect weather forecast data and observational solar intensity data for 10 months. The weather data includes temperature, dew point, wind speed, sky cover, probability of precipitation, and relative humidity. RMS-error of SVM-RBF is 27% more accurate than simple cloudy model and 51% more accurate than the PPF model. Dehalwar, Vasudev. et al used machine learning for efficient energy management in Sydney. Electricity load is observed for one week (24 hours) in Sydney region. They analyzed the relationship between electricity load and weather forecast using ANN and Bagged Regression Tree models and compared actual electric load and prediction models. They also compared the prediction accuracy ANN model with Bagged Regression Trees. The performance parameter is mean average error (MAE), the MAE of ANN and Bagged Regression Trees are 1.9% and 1.54% [5]. Furthermore, flow rate forecasting is used to calculate the inflow of water from a reservoir for the next few hours or

days in order to plan hydroelectric power scheduling and flood mitigation [6]. Meanwhile, weather data is used to forecast the daily stream flows. This paper uses four machine learning methods. Three of them are nonlinear models such as Bayesian neural network (BNN), Support Vector Regression (SVR) and Gaussian process (GP). The other one is linear model which is Multiple Linear Regression (MLR). Their performances were evaluated by comparing the observation value of stream flows and the machine learning forecast model. The result is that the nonlinear model generally perform better than the linear model), and the BNN had better scores (Corr and NSE, and lower MAE and RMSE) than other models.

Previous studies related to crude oil operation have actively researched the operation of crude oil for the maximization of profit. Furthermore, previous studies have optimized the management of crude oil inventory such as minimizing costs. These studies focused on how to plan to unload, mix, and feed. Lee et al. used MILP (Mixed Integer Linear Programming) to optimize unload crude oil with managing storage inventory [7]. Meanwhile, different crude oils have different components [8]. Hence, the product yield could be different when crude oil grades are blended. However, some constraints exist for CDUs such as sulfur and metals, because the sulfur specification of the product and metals acts as the permanent poison of a catalyst. Before charging crude oil to CDU, the predictive values of sulfur and metal should be calculated. It is related which kinds of crude oil are bought and decide the blending ratio of crude oil.

Furthermore, some studies focused on the transportation of crude oil. Crude oil could be transported by shipping and pipeline. Thus, several studies focused on shipping rather than pipeline because it has unsafely and uncertainty issues. Meanwhile, some studies focused on the safe transportation of crude oil. Hamzah studied and proposed the safety route for maritime [9]. Furthermore, Bouejla proposed a solution that prevents the crude oil theft by pirates, using Bayesian network [10]. Other studies focused on minimizing the cost of transportation of VLCC. As the price of crude oil changes and becomes uncertain, the product price also changes accordingly. This is a major problem in crude transportation worldwide. The world crude oil transport is the central transport that connects with the upstream and downstream sectors and plays an important role in world oil supply and demand market [11]. Cheng et al. formulated an optimization problem as a Markov decision process that integrates the uncertainties as transport time and crude oil demand [11]. For instance, it takes about 20 days for oil refineries in Korea to transport crude oil from Saudi Arabia. Owing to sea weather fluctuations, the storage tank capacity of crude oil has been fixed and it needs some time before feeding crude oil to CDU. Thus, it is very important to predict the time of arrival of the VLCC [12]. There are two problems in managing the schedule of VLCC. The VLCC may arrive early or late. If it arrives early, it has to wait to unload crude oil to the storage tank. Moreover, it needs to pay

waiting cost. On the other hand, if it arrives late, the throughput of downstream processes could be down. If the reserved oil is insufficient to be processed in a CDU at its minimum throughput, the late arrival of a vessel is disadvantageous because it will affect all the following refinery units [12].

Although there are many studies on crude oil unloading, blending, and charging, and transportation, no study has researched the relationship between meteorological data at a specific site and the unloading of crude oil from a VLCC. The same problem effects that VLCC arrives early or late can occur. For example, if it is predicted that VLCC can unload crude oil in a certain day, it may not be able to unload the crude oil due to sea weather changes. On the contrary, if it is predicted that VLCC cannot unload crude oil in a certain specific day, it may unload crude oil due to sea weather changes. In this case, the VLCC has to wait if there is not enough room to unload crude oil, thereby incurring waiting cost. The sea weather near the buoy is highly affected for VLCC to unload crude oil. Hence, this paper studies the relationship between sea weather and the unloading of crude oil by VLCC. Some studies have predicted sea weather in other fields.

A study related to wave forecast is being actively conducted in the field of ocean engineering. The numerical wave model has been widely used to forecast wave height. The sea state is described by a wave spectrum which is energy density. The numerical wave model predicts wave forecast solving wave spectrum using energy balance equation [13]. For instance, WAN (The WADMI Group, 1998), Wave watch III (Tolman, 1999) and SWAN (Booij et al, 1999) are famous numerical models. Recently, machine learning techniques are being used in the field of ocean engineering. Solomatine and Ostfeld studied data-driven modeling of water-related issues for river basin management [14]. Zijderveld dealt with the classification and prediction model of hydroponic issues using machine learning [15]. Corzo and Solomatine studied hydrological forecasting models using artificial neural networks (ANNs) [16]. Ahmadreza et al. improved the prediction accuracy of significant wave heights measured in few hours at the Caspian sea buoy. They made some wind and wave forecasts using machine learning algorithms such as ANN and Instance Based Learning (IBL), based on historical time series data. The prediction accuracy of IBL was found to be better than that of ANN for wave and wind condition one hour later [17].

A machine learning framework was developed to estimate ocean-wave conditions. James, Scott C., Zhang, and O'Donncha performed the supervised training of a machine learning model to predict the significant wave period and height [18]. These machine learning models were compared with SWAN (Simulating WAves Nearshore), which is the industry-standard wave-modeling tool [19]. The machine learning models (MLP, support vector machine, SVM) are accurate and computationally efficient surrogates for the SWAN model. Mahjoobi and Etemad-Shahidi proposed a prediction model that employs 5 years of wave and wind data gathered from Lake Michigan [20]. The input variables are

wind speed and wind direction, and the output variable is the significant wave heights. They used the regression tree, classification tree, decision tree, and ANN to build a classification model and compared the performances of the models. The error statics showed that the decision tree and ANN are similar and that the ANN is slightly more accurate than the classification tree and regression tree. Mahjoobi and Mosabbeb attempted to predict the wave height and wind using the SVM and ANNs. Compared to ANN, SVM has slightly better accuracy of SVM is slightly better, and the RBF kernel exhibits better performance than the polynomial kernel [21].

3. **Data and Methodology.**

3.1 Data

Three kinds of datasets are used. The first dataset is the time series of meteorological data for 3 years from January 1, 2016 to December 31, 2018. This dataset was obtained from the Meteorological Data Open Portal of the Korea Meteorological Administration (KMA; data.kma.go.kr). The dataset consists of 12 kinds of weather data, namely, wind speed, maximum wind speed, atmospheric pressure, humidity, air temperature, sea temperature, maximum wave height, significant wave height, average wave height, wave period, and wave direction. There are two kinds of wind speed data, three kinds of wave height, and seven different kinds of meteorological data. The second dataset comprises historical data on whether the VLCC unloads crude oil through single point mooring (SPM), i.e., the equipment that unloads crude oil from the vessel to the storage tank. These data are binary: if VLCC unloads crude oil, the value is 0; on the other hand, if the VLCC could not unload crude oil because of weather, the value is 1. This dataset was obtained from new.portmis.go.kr and referenced by an oil refinery company in the Republic of Korea. The data were obtained for the period from January 1, 2016 to December 31, 2018 and pertain to observations at Ulsan buoy. In this study, a model that predicts whether a VLCC unloads crude oil was built via SPM using the relations between weather observation data and the aforementioned binary data. Since future forecast is extremely important, the weather observation data are shifted from 1 h to 72 h later, and the relations are analyzed. For instance, in the case of 1 h later, the weather data value is 1/1 00:00, and the binary data value is 1/1 01:00. The relations of these two data are analyzed via machine learning techniques such as logistic regression, random forest, XGB and SVM. Thus, weather observation data are shifted from 1 h to 72 h sequentially. The last dataset comprises medium-term wave height forecast data from January 1, 2016 to December 31, 2018.This dataset is also obtained from the open portal of weather data of the KMA (data.kma.go.kr). This dataset is used for comparison to judge whether a VLCC unloads crude oil as the existing criteria: if the wave height exceeds 3 m, the VLCC cannot unload crude oil through SPM. This predictive accuracy is compared to the accuracy of machine learning.

3.1.1 Independent Variables.

The independent variables are meteorological data observed at Ulsan Buoy. The data analysis is based on relations between the weather observation data from January 1, 2016 to December 31, 2018 and on the historical time series data if VLCC unloads crude oil. The weather observation data consist of wind speed, maximum wind speed, atmospheric pressure, humidity, air temperature, sea temperature, maximum wave height, significant wave height, average wave height, wave period, and direction of wave, and these weather observation data are independent variables. The average, standard deviation, deviation, and box plots of independent variables are provided in this subsection. The weather forecast data represent maximum wind speed, direction of wind, air temperature, water temperature, significant wave height, and direction of waves on the internet such as www.kma.go.kr or www.windy.com. These types of data are referred when decision-makers judge whether VLCC unloads crude oil. So these six data types were used as independent variables. The average maximum wind speed is 8.7 m/s, and the upper quartile of the box plot, that is the top 25%, is above 20 m/s. Generally, the stronger wind speed, the more difficult it is to unload crude oil. This is because the tension of the rope that connects the VLCC and SPM is high. The average significant wave height is 1.2 m, and the upper quartile of box plot is 3 m. The higher the wave height, the more difficult it is to unload crude oil. This is because the tension of the rope that connects the VLCC and SPM is high. In experience, when the waves and wind are directed toward land rather than sea, it is difficult to unload crude oil.

Variables	Average	Standard deviation	Deviation
Wind speed (m/s)	6.4	3.2	10.4
Wind direction(deg)	176.2	111.0	12324.0
Max wind speed(m/s)	8.7	5.1	26.1
Atmospheric pressure(hPa)	1015.5	7.5	56.2
Humidity (%)	71.3	16.8	282.8
Air temperature(°C)	17.2	7.4	55.3
Sea temperature(°C)	19.8	4.6	21.4
Max wave height(m)	2.0	1.3	1.6
Significant wave height(m)	1.2	0.8	0.6
Average wave height(m)	0.9	0.6	0.3
Wave period(sec)	6.0	2.0	4.0
Wave direction(deg)	172.5	123.2	15169.8

Table 1. Average, standard deviation, and deviation values of independent variables.

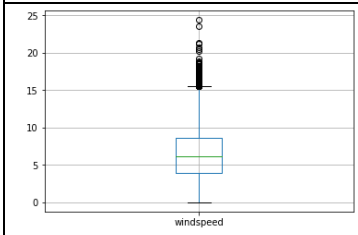
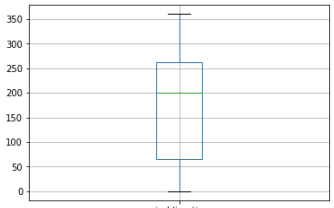
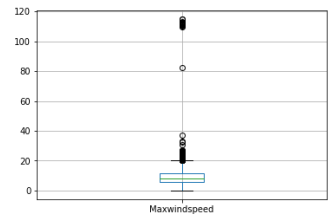
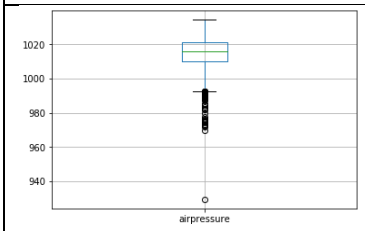
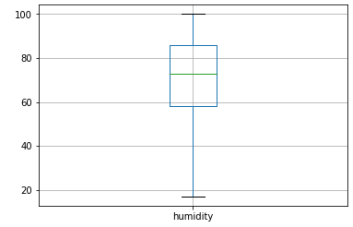
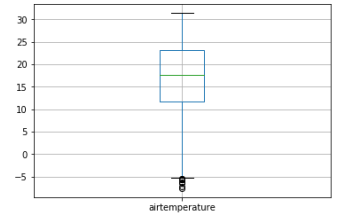
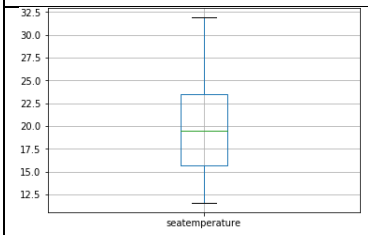
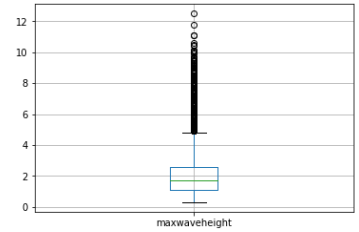
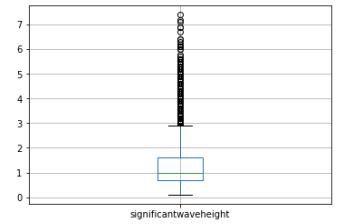
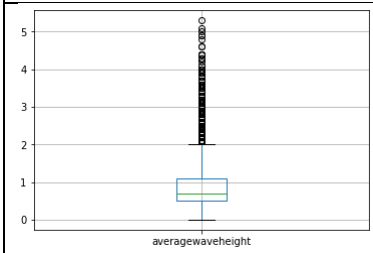
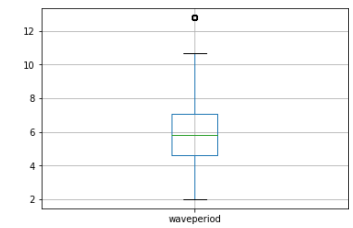
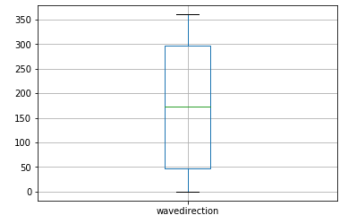
Wind speed (m/s)	Direction of wind(deg)	Max wind speed(m/s)
		
Atmospheric pressure(hPa)	**Humidity(%)**	**Air temperature(°C)**
		
Sea temperature(°C)	**Wind wave height(m)**	**Significant wave height(m)**
		
Average wave height(m)	**Wave period(sec)**	**Direction of wave(deg)**
		

Table 2. Boxplot of independent variables.

3.1.2 Variables Correlation

The same kinds of meteorological data such as maximum wind speed and wind speed have high positive correlation. The wind speed and maximum wind speed has positive correlation, and the value is 0.78. There are other similar kinds of meteorological data average wave height, significant wave height and maximum wave height, and they too have high positive correlation. The value of positive correlation between the average wave height and significant wave height is 1.0, and the value between the average wave height and maximum wave height is 0.97. The value of positive correlation between the significant wave height and maximum wave height is 0.97. The variables related to wave height and those related to wind have positive correlation. The value of positive correlation between wind speed and the variables related to wave height is from 0.64 to 0.65, and its value of the positive correlation between the maximum wind speed and the variables related to wave height is 0.57. The variables related to temperature such as air temperature and sea temperature have positive correlation, and the value is 0.86.

The atmospheric pressure has negative correlation with humidity and air temperature, and the values are -0.63 and -0.65, respectively. The air temperature has negative correlation with the variables related to wave heights, average wave height, significant wave height, and maximum wave height. The values of the correlation range from -0.37 to -0.35. In the case of sea temperature, the values of negative correlation with the variables related to wave heights range from -0.23 to -0.22.

	Wind Speed	Wind Direc	MaxWindSp	Air Pre.	Humidity	Air Temp.	Sea Temp.	MaxWave.H	Sig.Wave.H	Avg.Wave.H	Wave.Period	Wave.Direc
Wind Speed	1	0.08	0.79*	-0.07	-0.01	-0.18	-0.14	0.65*	0.64*	0.64*	0.06	0.05
Wind direction	0.08	1	0.09	0.12	-0.27*	-0.31*	-0.26*	-0.03	-0.04	-0.04	-0.02	-0.04
Max Wind Speed	0.79	0.09	1	0.01	-0.09	-0.24*	-0.16	0.57*	0.57*	0.57*	0.1	0.03
Air Pressure	-0.07	0.12	0.01	1	-0.63*	-0.65*	-0.5*	0.1	0.09	0.09	0.21	-0.13
Humidity	-0.01	-0.27	-0.09	-0.63*	1	0.62*	0.41	-0.18	-0.17	-0.17	-0.26*	0.17
Air Temperature	-0.18	-0.31	-0.24	-0.65*	0.62	1	0.87*	-0.37*	-0.35*	-0.35*	-0.29*	0.12
Sea Temperature	-0.14	-0.26	-0.16	-0.5*	0.41	0.87*	1	-0.23*	-0.22*	-0.22*	-0.13	0.04
Max Wave Height	0.65*	-0.03	0.57*	0.1	-0.18	-0.37*	-0.23*	1	0.97*	0.97*	0.52*	-0.02
Significant Wave Height	0.64*	-0.04	0.57*	0.09	-0.17	-0.35*	-0.22*	0.97*	1	1*	0.54*	-0.02
Average Wave Height	0.64*	-0.04	0.57*	0.09	-0.17	-0.35*	-0.22*	0.97*	1*	1	0.54*	-0.02
Wave Period	0.06	-0.02	0.1	0.21	-0.26*	-0.29*	-0.13	0.52*	0.54*	0.54*	1	-0.12
Wave Direction	0.05	-0.04	0.03	-0.13	0.17	0.12	0.04	-0.02	-0.02	-0.02	-0.12	1

Table 3. Correlation of independent variables.

3.1.3 Dependent Variables

There are 25,467 rows, and after excluding the rows that contain missing values, there are they remain 24,953 rows. Since the VLCC unloads the crude oil through the SPM, the column that indicates whether the VLCC unloads or not is added under the column name of SPM. The value is 0 if the VLCC unloads, and the value is 1 if the VLCC could not unload crude oil because of weather. In all, 4,722 rows represent that the VLCC could not unload because of weather, and these rows make up 19% of the total. It means that the proportion of cases in which the VLCC unloads the crude oil is 81%; therefore, under sampling, over sampling and SMOTE are used for data analysis.

3.2 Methodology

3.2.1 Logistic Regression.

The purpose of logistic regression is to present the relationship between dependent and independent variables as a specific function and use them in future predictive. Similar to linear regression, logistic regression describes linear coupling of independent variables to predict dependent variable. However, dependent variable of logistic regression targets categorical data and, when input data is given, the results of that data are divided into specific classifications. Logistic regression is used that dependent variable is binary. Logistic regression models are quite useful for classifying new cases into one of two outcome categories ("success" or "failure") [22]. Independent variables can be in any form, such as actual values, binary values, categories, etc. The form of dependent variables is divided into continuous (import, age, blood pressure) or discrete (sex, race). If more than one candidate with a specific discrete variable value exists, the candidates are generally converted into temporary variables to perform logistic regression. In other words, the separated independent variables are converted to have a value of '0' or '1' respectively. '0' means that the variable does not have a specific value, and '1' means that the variable is the same as the given value. Dependent variable Y is generally expressed as data from the Bernoulli distribution. Each dependent variable is determined by an unobserved probability p. This can be expressed in the following mathematical expressions:

$$Y_i | x_{1,i,\cdots,} x_{m,i} \ , Bernoulli(p_i)$$

3.2.2 Random Forest

A random forest is a tree-based ensemble with tree depending on a collection of random variables [23]. The method of random forest is the combinations of tree predictors. The decision tree focuses on the target variable is estimated by some values of input variables. It is a white box model and visual. The branches are divided by input variables, and it is connected to a child node. The decision tree is one of the typical supervised analyses. The characteristic of random forest is that the decision trees are formed randomly, and the decision trees have different characteristics. The typical method of random sampling of learning is bagging. The bagging forms different datasets of identical sizes, and learnings are proceeded by means of many different training data sets by allowing repetition using bootstrap. The random forest using bootstrap is that many training datasets are built, and the base learner is built after the training datasets are trained [24]. These base learners (decision trees) are combined by the method of average or majority voting and are made by a random forest. There are typical parameters such as *n* estimators and the maximum depth in random forest. The *n* estimators is the parameter that decide the number of model. The maximum depth is the parameter that decides the value of the maximum depth of the tree.

3.2.3 XGBoost

XGBoost is a framework for tree boosting. When the XGBoost makes a tree, a CART ensemble model is used. In case of CART, all leafs are related to the final score, unlike in the case of a decision tree. XGBoost also has extensibility in all scenarios and is 10 times faster than the other famous model in a single machine. The parallel and distributed computing greatly increases the speed. XGBoost is made by improving gradient-boosted and decision trees in terms of speed and performances [25]. The equation is as follows.

"Let y(t) be the prediction of the i-th instance at the t-th iteration. We will need to add f_t to minimize the following objective.

$$L^{(t)} \cong \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}^{(t-1)}\right) + g_t f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

After removing the constant terms, the following simplified objective at step t is obtained.

$$L^{(t)} \cong \sum_{i=1}^{n} \left[g_t f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

Rewrite the equation by expanding Ω as follows:

$$L^{(t)} = \sum_{i=1}^{n}\left[g_t f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2$$

$$= \sum_{i=1}^{n}\left[(\sum_{i\epsilon i_j} g_i)\omega_j + \frac{1}{2}(\sum_{i\epsilon i_j} h_i + \lambda)\omega_j^2\right] + \gamma T$$

The optimal weight ω_j can be calculated because q(X) is a fixed structure.

$$\omega_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} + \gamma T$$

The score function of tree q is as below.

$L^{(t)}(q) = -\frac{1}{2}\frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i+\lambda} + \gamma T$" [26]

XGBoost has three parameters, namely, the general parameter, booster parameter, and learning task parameter. The general parameters consist of the booster, n thread, and feature number. The booster parameter decides which structure is used, and the n-thread parameter decides the number of parameters. The feature number decides the number of dimensions. The maximum depth parameter is the one of the representative booster parameter, and it decides the maximum depth of the tree. If the number of the maximum depth parameter is high, it is easy to be over-fitted. The learning task parameter has the role of adjusting the objective function and the evaluation function of the model.

3.2.4. K-fold Cross validation

K-fold cross validation is equally partitioned into K-sized samples. One of the k samples is used to test the dataset, and the k-1 dataset is used as the training dataset, and this process is repeated k times. The average result after obtaining the value after k repetitions is used for the sake of accuracy. One dataset of k is used at least one time for validation. For instance, the five-fold validation is as follows.

1. Divide data sets S0, S1, S2, S3, and S4.
2. In case of the first validation, S0 is the validation set, and S1, S2, S3, and S4 are training sets.
3. In case of the second validation, S1 is the validation set, and S0, S2, S3, and S4 are the training sets.
4. In case of the third validation, S2 is the validation set, and S0, S1, S3, and S4 are the training sets.

5. In case of the fourth validation, S3 is the validation set, and S0, S1, S2, and S4 are the training sets.

6. In case of the fifth validation, S0 is the validation set, and S0, S1, S2, and S3 are the training sets.

3.2.5 ROC analysis

The ROC analysis is used to evaluate the prediction accuracy when dealing with the classification problem. The real data can be divided into P(positive) and N(negative), and the prediction data can be divided by Y(yes) and N(no). There are four cases, True Positive(TP), False Positive(FP), True Negative (TN), and False Negative (FN). The True Positive predicts positive, and the real data are positive. The False Positive predicts positive, and the real data are negative. The True negative predicts negative, and the real data are negative. The False negative predicts negative, and the real data are positive. The true positive and true negative predict well. It can be represented via a confusion matrix.

	Positive	Negative
Y	True Positive	False Positive
N	False Negative	True Negative

Some values are calculated as follows.

FP rate = $\frac{FP}{N}$

TP rate = $\frac{TP}{P}$

Precision=$\frac{TP}{TP+FP}$

Accuracy=$\frac{TP+TN}{P+N}$

recall = $\frac{TP}{P}$

The FP rate is X-axis, and TP rate is Y-axis on the ROC curve. If some points are above the line of ROC curve, they are called conservative. It means that true positive is bigger than false positive. If

some points are below the line, they are called liberal. It means that false negative is bigger than true positive [27].

3.2.6 Support Vector Machines

SVMs are used for the classification problem when the models are non-linear. The classification models are divided by the kernel, and the SVM is the algorithm that looks for boundaries with a big width. It makes the choice of the hyperplane that maximizes the distance of each class [28]. The SVM is divided by the shape of kernels. If the classes are divided by a line, the kernel is linear. If the classes cannot be divided by a line, the kernel can be decided by RBF, polynomial, and sigmoid. There are parameters such as cost and gamma. The parameter cost decides about allowing how many samples should be put in other classes. If the parameter costs were larger, lower samples can be assigned in other classes. If the parameter costs were lower, larger samples can be assigned in other classes. The parameter gamma shows the influence of the sample. If gamma is large, the standard deviation is low, and it makes the shape of the kernel more winding. If gamma is lower, the standard deviation is larger, and it makes the shape of kernel more fluent. It is important for the performance to set these two parameters appropriately.

4. Result

4.1 Predictive accuracy of unloading crude oil with the existing criteria

The mid-term sea weather forecast at Ulsan buoy from January 1, 2016 to December 31, 2018 is used, and the gap between the forecast date and release date is from 2.5 days to 10 days. The existing criteria is that the VLCC could not unload crude oil exceed the significant wave height of 3 m. Hence, the real case whether or not the VLCC unloads the crude oil and the prediction case based on mid-term sea weather forecast are compared. The predictive accuracy is 75%, and the accuracy is not different at the gap between the forecast date and release date. There are 27,488 cases. The cases of mistaken prediction are important. One mistaken predictions is one that predicts "open" (the VLCC unloads crude oil) but gives the result "close" (the VLCC could not unload crude oil because of weather). The other mistaken prediction is one that predicts "close" but gives the result "open." The number of the first mistaken prediction case is 2,607 and the miss rate ($\frac{FN}{P}$) is 13%. The number of the second mistaken prediction case is 4,329. The fall-out rate ($\frac{FP}{N}$) is 60%.

Predict. / Real	Open	Close
Open	17719	4329
Close	2607	2833
	13%	60%

Table 4. Confusion matrix of prediction using the existing criteria.

Gap between release and forecast (days)	Accuracy (%)
2.5	79
3	78
3.5	77
4.0	77
4.5	76
5.0	75
5.5	75
6.0	73
6.5	73
7.0	72
7.5	72
8.0	73
8.5	73
9.0	73
9.5	75
10	75

Table 5. Predictive accuracy according to the gap between release and forecasts.

4.2 Prediction of unloading crude oil using machine learning

The classification model analyzes the relations between the binary data that indicates if the VLCC unloads crude oil and the observed meteorological data. The meteorological variables obtained on the internet provided by the KMA (www.kma.go.kr) or Windy (www.windy.co.kr) are used. The variables consist of significant wave height, wave direction, air temperature, sea temperature, max wind speed, and wind direction, and so these variables are used as independent variables in machine learning. The prediction in future is more important than that at present, and so machine learning is proceeded such that rows of meteorological data are staggered. For instance, meteorological data are 1/1 00:00 and the history if the VLCC unloads the crude oil is 1/1 1:00; that is, the meteorological data are staggered for 1 h. These processes are performed for 1 h, 6 h, 12 h, 24 h, 36 h, 48 h, 60 h, and 72 h. The methods of machine learning are used as logistic regression, random forest, XGBoost, and SVM. The case that the VLCC unloads the crude oil accounts for 81%, so under sampling, over sampling and SMOTE are used, and the dataset is divided into the training set and test set in the ratio 7:3.

4.2.1 Predict if the VLCC unloads crude oil at time t using meteorological data at time t

First, the analysis is performed to predict if the VLCC unloads crude oil at time t using meteorological data at time t. The predictive accuracy of random forest and XGBoost is better than that of the logistic regression and RBF SVM. The predictive accuracy of random forest and XGBoost is 98%, and that of logistic regression and RBR SVM are 89% and 92%, respectively. The random forest and XGBoost with over sampling have the best performance (98%). The AUC score of the random forest and XGBoost with over sampling is also the best (98%). The predictive accuracy of cross validation about XGBoost is the best at 95%. The miss rate ($\frac{FN}{P}$) of the predictive accuracy existing criteria is 13%, and the fall-out rate ($\frac{FP}{N}$) of the predictive accuracy existing criteria is 60%. They are compared to check if the performance of the predictive model is better. Based on the result of confusion matrix, the logistic regression with basic sampling has better performance than comparison, but other logistic regression models with under sampling, oversampling and SMOTE have slightly lower performance (high miss rate) (17%, 16% and 15%) than comparison (13%). The miss rate of SVM with under sampling has also lower performance (19%) than comparison (13%). But the random forest and XGBoost have better performances (low miss rate and fall out rate) than comparison regardless of the sampling.

		Accuracy	AUC	5-Fold	10-Fold
LR	Sample	89%	78%	89%	89%
	Over	85%	85%		
	SMOTE	86%	86%		
	Under	84%	84%		
Random Forest	Sample	93%	85%	94%	94%
	Over	98%	98%		
	SMOTE	95%	95%		
	Under	88%	88%		
XGB	Sample	95%	89%	96%	95%
	Over	98%	98%		
	SMOTE	97%	97%		
	Under	90%	90%		
SVM	Sample	90%	79%	90%	90%
	Over	91%	91%		
	SMOTE	92%	92%		
	Under	81%	81%		

Table 6. Predictive accuracy in case of 4.2.1

	Sampling			Oversampling			SMOTE			Undersampling		
Logistic regression	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	5948	233	Open	5327	806	Open	5327	806	Open	1254	185
	Close	579	832	Close	982	5170	Close	913	5239	Close	265	1194
		9%	22%		16%	13%		15%	13%		17%	13%
Random Forest	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	6042	139	Open	5880	253	Open	5791	342	Open	1299	140
	Close	379	1032	Close	46	6106	Close	265	5887	Close	198	1261
		6%	12%		1%	4%		4%	5%		13%	10%
XGBoost	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	6045	136	Open	5921	212	Open	5944	189	Open	1280	159
	Close	274	1137	Close	42	6110	Close	226	5926	Close	134	1325
		4%	11%		1%	3%		4%	3%		9%	11%
SVM	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	5973	208	Open	5431	702	Open	5477	656	Open	1153	286
	Close	552	859	Close	358	5794	Close	354	5798	Close	275	1184
		8%	19%		6%	11%		6%	10%		19%	19%

Table 7. Confusion matrix in case of 4.2.1

4.2.2 Predict if the VLCC unloads crude oil at time t using meteorological data at time t-1 hour

Second, an analysis is performed to predict if the VLCC unloads crude oil at time t using meteorological data 1 h before time t. The result is almost identical as that of the first analysis. The predictive accuracy of random forest and XGBoost is better than that of logistic regression and RBF SVM. The predictive accuracy of random forest and XGBoost is 98%, and that of logistic regression and RBF SVM are 89% and 92%, respectively. The random forest and XGBoost with over sampling have the best performance (98%). The AUC score of random forest and XGBoost with over sampling has also the best performance (98%). The predictive accuracy of cross validation about XGBoost is the best at 94%. The miss rate ($\frac{FN}{P}$) of the predictive accuracy existing criteria is 13%, and the fall-out rate ($\frac{FP}{N}$) of the predictive accuracy of the existing criteria is 60%. They are compared to check whether the performance of the predictive model is better. Based on the result of the confusion matrix, the logistic regression with basic sampling has better performance than comparison, but other logistic regression models with under sampling, over sampling and SMOTE have slightly lower performance (17%, 16% and 14%, respectively) than comparison (13%). The miss rate of SVM with under sampling has also lower performance (18%) than comparison (13%). But the random forest and XGBoost have better performances (low miss rate and fall out rate) than comparison regardless of the sampling.

		Accuracy	AUC	5-Fold	10-Fold
LR	Sample	89%	78%	89%	89%
	Over	85%	85%		
	SMOTE	86%	86%		
	Under	85%	85%		
Random Forest	Sample	93%	86%	94%	94%
	Over	98%	98%		
	SMOTE	95%	95%		
	Under	90%	90%		
XGB	Sample	94%	89%	94%	95%
	Over	98%	98%		
	SMOTE	96%	96%		
	Under	91%	91%		
SVM	Sample	89%	80%	90%	90%
	Over	92%	92%		
	SMOTE	92%	92%		
	Under	82%	82%		

Table 8. The predictive accuracy in case of 4.2.2

	Sampling			Oversampling			SMOTE			Under		
Logistic regression	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	5932	249	Open	5337	796	Open	5337	796	Open	1268	171
	Close	564	847	Close	1008	5144	Close	846	5306	Close	252	1207
		9%	23%		16%	13%		14%	13%		17%	12%
Random Forest	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	6040	141	Open	5892	241	Open	5765	368	Open	1287	152
	Close	357	1054	Close	54	6098	Close	234	5918	Close	152	1307
		6%	12%		1%	4%		4%	6%		11%	10%
XGBoost	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	6043	138	Open	5923	210	Open	5922	211	Open	1309	130
	Close	292	1119	Close	42	6110	Close	256	5896	Close	129	1330
		5%	11%		1%	3%		4%	3%		9%	9%
SVM	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	5843	338	Open	5457	676	Open	5472	661	Open	1195	244
	Close	474	937	Close	365	5787	Close	337	5815	Close	271	1188
		8%	27%		6%	10%		6%	10%		18%	17%

Table 9. Confusion matrix in case of 4.2.2

4.2.3 Predict if the VLCC unloads crude oil at time t using meteorological data at time t-6 hour

Third, the analysis is performed to predict if the VLCC unloads crude oil at time t hour using meteorological data 6 h before time t. The result is almost same as that of the first one. The predictive accuracy of random forest and XGBoost is better than that of the logistic regression and RBF SVM. The predictive accuracy of random forest and XGBoost is 98%, and the predictive accuracy of logistic regression is 89% and that of RVF SVM is 91%. The random forest and XGBoost with over sampling have the best performance (98%). The AUC score of random forest and XGBoost with over sampling has the best performance (98%). The predictive accuracy of cross validation about XGBoost is the best at 95%. The miss rate ($\frac{FN}{P}$) of the predictive accuracy existing criteria is 13%, and the fall-out rate ($\frac{FP}{N}$) of the predictive accuracy existing criteria is 60%. They are compared to check whether the performance of the predictive model is better. Based on the result of confusion matrix, the miss rate of the logistic regression with basic sampling has better performance than the comparison, but others with under sampling, oversampling and SMOTE have slightly lower performance (17%, 17% and 15% respectively) than the comparison (13%). The miss rate of SVM with under sampling has also lower performance (19%) than comparison (13%). But the random forest and XGBoost have better performances (low miss rate and fall out rate) than comparison regardless of the sampling.

		Accuracy	AUC	5-Fold	10-Fold
LR	Sample	89%	76%	89%	89%
	Over	84%	84%		
	SMOTE	85%	85%		
	Under	84%	84%		
Random Forest	Sample	93%	85%	94%	94%
	Over	98%	98%		
	SMOTE	95%	95%		
	Under	89%	89%		
XGB	Sample	94%	87%	94%	95%
	Over	98%	98%		
	SMOTE	96%	96%		
	Under	91%	91%		
SVM	Sample	89%	80%	90%	90%
	Over	91%	91%		
	SMOTE	91%	91%		
	Under	81%	81%		

Table 10. Predictive accuracy in case of 4.2.3

	Sampling			Oversampling			SMOTE			Under		
Logistic regression	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	5941	241	Open	5277	698	Open	5277	868	Open	1244	209
	Close	612	793	Close	1109	5033	Close	934	5208	Close	263	1181
		9%	23%		17%	12%		15%	14%		17%	15%
Random Forest	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	6041	141	Open	5905	240	Open	5805	340	Open	1304	149
	Close	385	1025	Close	36	6106	Close	304	5838	Close	161	1283
		6%	12%		1%	4%		5%	6%		11%	10%
XGBoost	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	6040	142	Open	5923	222	Open	5952	193	Open	1319	134
	Close	288	1122	Close	32	6110	Close	253	5889	Close	139	1305
		5%	11%		1%	4%		4%	3%		10%	9%
SVM	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	5830	352	Open	5430	715	Open	5479	666	Open	1179	274
	Close	489	921	Close	387	5755	Close	384	5758	Close	277	1167
		8%	28%		7%	11%		7%	10%		19%	19%

Table 11. Confusion matrix in case of 4.2.3

4.2.4 Predict if the VLCC unloads crude oil at time t using meteorological data at time t-12 hour

Fourth, the analysis that predicts if the VLCC unloads crude oil at time t hour using meteorological data 12 h before time t is performed. The predictive accuracy of XGBoost and random forest is almost the same as that of the first one. The predictive accuracy of random forest and XGBoost is better than that of logistic regression and SVM. The predictive accuracy of random forest and XGBoost is 98%, and that of logistic is 87% and that of SVM is 89%. The random forest and XGBoost with over sampling have the best performance (98%). The AUC score of random forest and XGBoost with over sampling is also the best (98%). The predictive accuracy of cross validation about XGBoost is the best at 94%. The miss rate ($\frac{FN}{P}$) of the predictive accuracy existing criteria is 13%, and the fall-out rate ($\frac{FP}{N}$) of the predictive accuracy existing criteria is 60%. They are compared to check whether the performance of predictive model is better. Based on the result of the confusion matrix, the miss rate of logistic regression with basic sampling has better performance than the comparison, but others with under sampling, oversampling and SMOTE have lower performance (22%, 21% and 20%, respectively) than the comparison (13%). The miss rate of random forest and SVM with under sampling has lower performance (15% and 22%) than comparison (13%). But XGBoost has better performances (low miss rate and fall out rate) than comparison regardless of the sampling.

		Accuracy	AUC	5-Fold	10-Fold
LR	Sample	87%	72%	88%	88%
	Over	80%	80%		
	SMOTE	82%	82%		
	Under	80%	80%		
Random Forest	Sample	92%	83%	93%	93%
	Over	98%	98%		
	SMOTE	94%	94%		
	Under	86%	86%		
XGB	Sample	93%	87%	94%	94%
	Over	98%	98%		
	SMOTE	96%	96%		
	Under	88%	88%		
SVM	Sample	89%	76%	89%	89%
	Over	89%	89%		
	SMOTE	89%	89%		
	Under	79%	79%		

Table 12.Predictive accuracy in case of 4.2.4

	Sampling			Oversampling			SMOTE			Under		
Logistic regression	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	5935	219	Open	5152	1012	Open	5152	1012	Open	1195	239
	Close	742	692	Close	1406	4711	Close	1256	4861	Close	344	1117
		11%	24%		21%	18%		20%	17%		22%	18%
Random Forest	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	6014	140	Open	5921	243	Open	5777	387	Open	1272	162
	Close	468	966	Close	49	6068	Close	364	5753	Close	231	1230
		7%	13%		1%	4%		6%	6%		15%	12%
XGBoost	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	5988	166	Open	5964	200	Open	5960	204	Open	1279	155
	Close	342	1092	Close	45	6072	Close	315	5802	Close	186	1275
		5%	13%		1%	3%		5%	3%		13%	11%
SVM	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	5912	242	Open	5356	808	Open	5383	781	Open	1134	300
	Close	622	812	Close	549	5568	Close	554	5563	Close	314	1147
		10%	23%		9%	13%		9%	12%		22%	21%

Table 13. Confusion matrix in case of 4.2.4

4.2.5 Predict if the VLCC unloads crude oil at time t using meteorological data at time t-24 hour

Fifth, the analysis that predicts if the VLCC unloads crude oil at time t hour using meteorological data 24 h before time t is performed. The predictive accuracy of XGBoost and random forest is slightly lower than that of the first one. The predictive accuracy of random forest and XGBoost is better than that of logistic regression and RBF SVM. The predictive accuracy of random forest and XGBoost is 97%, and that of logistic regression is 84% and that of RBF SVM is 85%. The random forest and XGBoost with over sampling have the best performance (97%). The AUC score of random forest and XGBoost with over sampling is also the best (97%). The predictive accuracy of cross validation about XGBoost is the best at 92%. The miss rate ($\frac{FN}{P}$) of the predictive accuracy existing criteria is 13%, and the fall-out rate ($\frac{FP}{N}$) of the predictive accuracy existing criteria is 60%. They are compared to check whether the performance of the predictive model is better. Based on the result of the confusion matrix, the miss rate of logistic regression models regardless of sampling have lower performance (15%, 31%, 30% and 31% respectively) than comparison (13%). The miss rate of the SVM models with oversampling and SMOTE have lower performance (15% and 14%, respectively) than comparison (13%), but the SVM model with basic sampling has slightly better performance than the comparison. The miss rate of all models with under sampling has lower performance (31%, 16%, 14% and 28%) than comparison (13%). But the random forest and XGBoost with over sampling and SMOTE have better performances (low miss rate and fall out rate) than comparison.

		Accuracy	AUC	5-Fold	10-Fold
LR	Sample	84%	61%	84%	84%
	Over	71%	71%		
	SMOTE	72%	72%		
	Under	72%	72%		
Random Forest	Sample	90%	76%	90%	90%
	Over	97%	97%		
	SMOTE	92%	92%		
	Under	85%	85%		
XGB	Sample	92%	84%	92%	92%
	Over	97%	97%		
	SMOTE	95%	95%		
	Under	86%	86%		
SVM	Sample	85%	70%	86%	86%
	Over	84%	84%		
	SMOTE	84%	84%		
	Under	73%	73%		

Table 14. Predictive accuracy in case of 4.2.5

	Sampling			Oversampling			SMOTE			Under		
Logistic regression	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	6035	137	Open	4705	1420	Open	4705	1420	Open	1095	339
	Close	1075	338	Close	2157	3991	Close	2030	4118	Close	482	980
		15%	29%		31%	26%		30%	26%		31%	26%
Random Forest	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	6043	129	Open	5876	249	Open	5702	423	Open	1246	188
	Close	644	769	Close	104	6044	Close	529	5619	Close	235	1227
		10%	14%		2%	4%		8%	7%		16%	13%
XGBoost	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	6016	156	Open	5875	250	Open	5904	221	Open	1238	196
	Close	423	990	Close	79	6069	Close	378	5770	Close	202	1260
		7%	14%		1%	4%		6%	4%		14%	13%
SVM	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	5855	317	Open	5004	1121	Open	5014	1111	Open	1059	375
	Close	784	629	Close	897	5251	Close	814	5334	Close	405	1057
		12%	34%		15%	18%		14%	17%		28%	26%

Table 15. Confusion matrix in case of 4.2.5

4.2.6 Predict if the VLCC unloads crude oil at time t using meteorological data at time t-36 hour

Sixth, the analysis is conducted to predict if the VLCC unloads crude oil at time t hour using meteorological data 36 h before time t. The predictive accuracy of XGBoost and random forest is slightly lower than that of the first one. The predictive accuracy of random forest and XGBoost is better than that of logistic regression and RBF SVM. The predictive accuracy of random forest and XGBoost are 98% and 97% respectively, and that of logistic regression is 81% and that of RBF SVM is 83%. The random forest with over sampling have the best performance (98%). The AUC score of random forest with over sampling is also the best (98%). The predictive accuracy of cross validation about XGBoost is the best at 92%. The miss rate ($\frac{FN}{P}$) of the predictive accuracy existing criteria is 13%, and the fall-out rate ($\frac{FP}{N}$) of the predictive accuracy existing criteria is 60%. They are compared to check whether the performance of the predictive model is better. Based on the result of the confusion matrix, the miss rate of logistic regression models regardless of sampling have lower performance (18%, 40%, 39% and 38% respectively) than comparison (13%). The miss rate of the SVM model has lower performance (15%, 18%, 19% and 36%) than comparison (13%) regardless sampling. The miss rate of all models with under sampling has lower performance (38%, 17%, 15% and 36%) than comparison (13%). But the random forest and XGBoost with over sampling and SMOTE have better performances (low miss rate and fall out rate) than comparison.

		Accuracy	AUC	5-Fold	10-Fold
LR	Sample	81%	53%	81%	82%
	Over	63%	63%		
	SMOTE	63%	63%		
	Under	63%	63%		
Random Forest	Sample	89%	73%	89%	90%
	Over	98%	98%		
	SMOTE	92%	92%		
	Under	84%	84%		
XGB	Sample	92%	83%	92%	92%
	Over	97%	97%		
	SMOTE	95%	95%		
	Under	85%	85%		
SVM	Sample	83%	63%	84%	84%
	Over	79%	79%		
	SMOTE	80%	80%		
	Under	64%	64%		

Table 16. Predictive accuracy in case of 4.2.6

	Sampling			Oversampling			SMOTE			Under		
Logistic regression	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	6070	62	Open	4345	1751	Open	4345	1751	Open	1017	426
	Close	1357	92	Close	2839	3339	Close	2809	3368	Close	629	817
		18%	40%		40%	34%		39%	34%		38%	34%
Random Forest	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	6057	75	Open	5872	224	Open	5659	437	Open	1216	227
	Close	759	690	Close	72	6105	Close	592	5585	Close	248	1198
		11%	10%		1%	4%		9%	7%		17%	16%
XGBoost	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	5975	157	Open	5806	290	Open	5852	244	Open	1211	232
	Close	470	979	Close	41	6136	Close	430	5747	Close	215	1231
		7%	14%		1%	5%		7%	4%		15%	16%
SVM	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	5798	334	Open	4979	1417	Open	4665	1431	Open	934	509
	Close	987	462	Close	1100	5077	Close	1075	5102	Close	518	928
		15%	42%		18%	22%		19%	22%		36%	35%

Table 17. Predictive accuracy in case of 4.2.6

4.2.7 Predict if the VLCC unloads crude oil at time t using meteorological data at time t-48 h (2 days)

The sixth analysis is conducted to predict if the VLCC unloads crude oil at time t hour using meteorological data 48 h (2 days) before time t. The predictive accuracy of XGBoost and random forest is slightly lower than first one. The predictive accuracy of random forest and XGBoost is better than that of logistic regression and RBF SVM. The predictive accuracy of random forest and XGBoost is 97% and that of logistic regression is 81% and that of RBF SVM is 81%. The random forest and XGBoost with over sampling have the best performance (97%). The AUC score of random forest and XGBoost with over sampling is also the best (97%). The predictive accuracy of cross validation about XGBoost is the best at 91%. The miss rate ($\frac{FN}{P}$) of the predictive accuracy existing criteria is 13%, and the fall-out rate ($\frac{FP}{N}$) of the predictive accuracy existing criteria is 60%. They are used for comparison to check whether the performance of the predictive model is better. Based on the result of the confusion matrix, the miss rate of logistic regression models regardless of sampling have lower performance (19%, 43%, 43% and 42% respectively) than comparison (13%). The miss rate of the SVM model has lower performance (16%, 21%, 19% and 37%) than comparison (13%) regardless sampling. The miss rate of all models with under sampling has lower performance (42%, 15%, 13% and 37%) than comparison (13%). (The miss rate of XGBoost with under sampling is same as comparison). But the random forest and XGBoost with over sampling and SMOTE have better performances (low miss rate and fall out rate) than comparison.

		Accuracy	AUC	5-Fold	10-Fold
LR	Sample	81%	51%	81%	82%
	Over	58%	58%		
	SMOTE	58%	58%		
	Under	58%	58%		
Random Forest	Sample	81%	71%	89%	90%
	Over	97%	97%		
	SMOTE	92%	92%		
	Under	83%	83%		
XGB	Sample	94%	80%	91%	91%
	Over	97%	97%		
	SMOTE	94%	94%		
	Under	85%	85%		
SVM	Sample	81%	58%	82%	82%
	Over	77%	77%		
	SMOTE	78%	78%		
	Under	63%	63%		

Table 18. Predictive accuracy in case of 4.2.7

	Sampling			Oversampling			SMOTE			Under		
Logistic regression	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	6111	25	Open	3884	2203	Open	3884	2203	Open	909	551
	Close	1419	23	Close	2961	3213	Close	2986	3188	Close	659	775
		19%	52%		43%	41%		43%	41%		42%	42%
Random Forest	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	6058	78	Open	5816	271	Open	5577	510	Open	1181	279
	Close	827	615	Close	82	6092	Close	521	5653	Close	203	1231
		12%	11%		1%	4%		9%	8%		15%	18%
XGBoost	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	5958	178	Open	5783	304	Open	5841	246	Open	1200	260
	Close	523	919	Close	75	6099	Close	473	5701	Close	187	1247
		8%	16%		1%	5%		7%	4%		13%	17%
SVM	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	5808	328	Open	4440	1647	Open	4396	1691	Open	929	531
	Close	1123	319	Close	1153	5021	Close	1018	5156	Close	548	886
		16%	51%		21%	25%		19%	25%		37%	37%

Table 19. Confusion matrix in case of 4.2.7

4.2.8. Predict if the VLCC unloads crude oil at time t using meteorological data at time t-60 h (2.5 days)

The eighth analysis is performed to predict if the VLCC unloads crude oil at time t hour using meteorological data 60 h (2.5 days) before time t. The predictive accuracy of XGBoost and random forest is slightly lower than that of the first one. The predictive accuracy of random forest and XGBoost is better than that of the logistic regression and RBF SVM. The predictive accuracy of random forest and XGBoost is 97%, and that of logistic regression is 81%, and that of RBF SVM is 80%. The random forest and XGBoost with over sampling have the best performance (97%). The AUC score of random forest and XGBoost with over sampling is also the best (97%). The predictive accuracy of cross validation for XGBoost is the best at 91%. The miss rate ($\frac{FN}{P}$) of the predictive accuracy existing criteria is 13%, and the fall-out rate ($\frac{FP}{N}$) of the predictive accuracy with the existing criteria is 60%. Comparisons are made to check whether the performance of the predictive model is better. Based on the result of the confusion matrix, the miss rate of logistic regression models regardless of sampling have lower performance (19%, 45%, 45% and 44% respectively) than comparison (13%). The fall out rate of logistic regression with sampling has lower performance (75%) than comparison (60%). The miss rate of the SVM model has lower performance (17%, 21%, 20% and 39%) than comparison (13%) regardless sampling. The miss rate of all models with under sampling has lower performance (44%, 17%, 17% and 39%) than comparison (13%). But the random forest and XGBoost with oversampling and SMOTE have better performances (low miss rate and fall out rate) than comparison.

		Accuracy	AUC	5-Fold	10-Fold
LR	Sample	81%	50%	81%	82%
	Over	56%	56%		
	SMOTE	56%	56%		
	Under	56%	56%		
Random Forest	Sample	88%	70%	88%	89%
	Over	97%	97%		
	SMOTE	91%	91%		
	Under	82%	82%		
XGB	Sample	91%	79%	91%	91%
	Over	97%	97%		
	SMOTE	94%	94%		
	Under	82%	82%		
SVM	Sample	80%	57%	81%	82%
	Over	76%	76%		
	SMOTE	77%	77%		
	Under	61%	61%		

Table 20. The predictive accuracy in case of 4.2.8

	Sampling			Oversampling			SMOTE			Under		
Logistic regression	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	6104	3	Open	3374	2704	Open	3836	2242	Open	766	668
	Close	1466	1	Close	2727	3447	Close	3180	2994	Close	603	859
		19%	75%		45%	44%		45%	43%		44%	44%
Random Forest	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	6052	55	Open	5849	229	Open	5560	518	Open	1147	287
	Close	871	596	Close	82	6092	Close	600	5574	Close	231	1231
		13%	8%		1%	4%		10%	9%		17%	19%
XGBoost	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	5980	127	Open	5779	299	Open	5838	240	Open	1147	287
	Close	589	878	Close	79	6095	Close	503	5671	Close	236	1226
		9%	13%		1%	5%		8%	4%		17%	19%
SVM	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	5777	330	Open	4335	1743	Open	4318	1760	Open	872	563
	Close	1179	288	Close	1162	5012	Close	1072	5102	Close	563	899
		17%	53%		21%	26%		20%	26%		39%	39%

Table 21. Confusion matrix in case of 4.2.8

4.2.9. Predict if the VLCC unloads crude oil at time t using meteorological data at time t-72 h (3 days).

Finally, analysis is conducted to predict if the VLCC unloads crude oil at time t hour using meteorological data 72 h (3 days) before time t. The predictive accuracy of XGBoost and random forest is slightly lower than that of the first one. The predictive accuracy of random forest and XGBoost is better than that of logistic regression and RBF SVM. The predictive accuracy of random forest and XGBoost is 97%, and that of logistic regression is 80% and that of RBF SVM is 80%. The random forest and XGBoost with over sampling have the best performance (97%). The AUC score of random forest and XgBoost with over sampling is also the best (97%). The predictive accuracy of cross validation for XGBoost is the best at 91%. The miss rate ($\frac{FN}{P}$) of the predictive accuracy of the existing criteria is 13%, and the fall-out rate ($\frac{FP}{N}$) of the predictive accuracy of the existing criteria is 60%. They are compared to decide whether the performance of the predictive model is better. Based on the result of the confusion matrix, the miss rate of logistic regression models regardless of sampling have lower performance (20%, 45%, 45% and 45% respectively) than comparison (13%). The fall out rate of logistic regression with sampling has lower performance (100%) than comparison (60%). The miss rate of the SVM model has lower performance (18%, 22%, 19% and 39%) than comparison (13%) regardless sampling. The miss rate of all models with under sampling has lower performance (45%, 15%, 15% and 39%) than comparison (13%). But the random forest and XGBoost with oversampling and SMOTE have better performances (low miss rate and fall out rate) than comparison.

		Accuracy	AUC	5-Fold	10-Fold
LR	Sample	80%	50%	81%	81%
	Over	55%	55%		
	SMOTE	55%	55%		
	Under	54%	54%		
Random Forest	Sample	87%	68%	88%	88%
	Over	97%	97%		
	SMOTE	90%	90%		
	Under	82%	82%		
XGB	Sample	89%	77%	91%	91%
	Over	97%	97%		
	SMOTE	94%	94%		
	Under	83%	83%		
SVM	Sample	80%	55%	81%	81%
	Over	75%	75%		
	SMOTE	76%	76%		
	Under	60%	60%		

Table 22. Predictive accuracy in case of 4.2.9

	Sampling			Oversampling			SMOTE			Under		
Logistic regression	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	6089	2	Open	3055	3057	Open	3055	3057	Open	698	758
	Close	1479	0	Close	2513	3624	Close	2514	3623	Close	580	856
		20%	100%		45%	46%		45%	46%		45%	47%
Random Forest	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	6033	58	Open	5885	227	Open	5458	654	Open	1140	316
	Close	925	554	Close	94	6043	Close	520	5617	Close	207	1229
		13%	9%		2%	4%		9%	10%		15%	20%
XGBoost	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	5923	168	Open	5797	315	Open	5894	218	Open	1164	292
	Close	631	848	Close	67	6070	Close	507	5630	Close	201	1235
		10%	17%		1%	5%		8%	4%		15%	19%
SVM	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close	Predict. / Real	Open	Close
	Open	5798	293	Open	4230	1882	Open	4242	1870	Open	848	608
	Close	1247	232	Close	1225	4912	Close	1019	5118	Close	551	885
		18%	56%		22%	28%		19%	27%		39%	41%

Table 23. Confusion matrix in case of 4.2.9

4.3 Variable importance.

The random forest and XGBoost with over sampling have the highest predictive accuracy and best performance (low miss rate and fall out rate). The most important variable for the random forest analysis of the current prediction is significant wave height, which is 0.43. The value of it decreases as we predict a further future. The value of variable importance with significant wave height of 72 hours after the current decreases 0.12. On the contrary, other variables except significant wave height and max wind speed increase as we predict a further future. Especially, the value of variable importance with sea temperature of 72 hours after the current has risen to 0.24.

The most important variable for the XGBoost analysis of the current prediction is significant wave height, which is 0.75. The value of it decreases as we predict a further future. The value of variable importance with significant wave height of 72 hours after the current has decreased to 0.2. On the contrary, other variables except significant wave height and max wind speed show increasing tendency as we predict a further future. Especially, the value of variable importance with sea temperature of 72 hours after the current has risen to 0.23.

		variable importance					
		wind direction	Max wind speed	air temperature	sea temperature	significant wave height	wave direction
Random Forest	0h	0.11	0.16	0.12	0.12	0.42	0.07
	1h	0.11	0.16	0.11	0.12	0.42	0.07
	6h	0.12	0.19	0.12	0.12	0.38	0.07
	12h	0.14	0.2	0.13	0.14	0.3	0.1
	24h	0.16	0.2	0.17	0.18	0.17	0.12
	36h	0.16	0.15	0.2	0.21	0.14	0.13
	48h	0.16	0.14	0.2	0.23	0.13	0.14
	60h	0.16	0.14	0.2	0.24	0.12	0.14
	72h	0.16	0.14	0.2	0.24	0.12	0.14
XGB	0h	0.05	0.04	0.07	0.06	0.75	0.04
	1h	0.05	0.04	0.07	0.06	0.75	0.04
	6h	0.07	0.06	0.08	0.07	0.69	0.04
	12h	0.09	0.09	0.1	0.1	0.56	0.06
	24h	0.16	0.26	0.15	0.16	0.18	0.1
	36h	0.13	0.14	0.17	0.21	0.24	0.11
	48h	0.14	0.13	0.19	0.22	0.2	0.12
	60h	0.13	0.14	0.18	0.24	0.19	0.13
	72h	0.14	0.13	0.18	0.23	0.2	0.12

Table 24. Variable importance.

5. Conclusion

This study has theoretical contribution that showed the usefulness of machine learning for specific purpose in local area with weather data. This study has also improved the predictive accuracy of works if the VLCC (Very Large Crude Carriers) unloads crude oil to onshore crude storage tank based on weather forecast. Before this study, the predictive accuracy at SPM (Single Point Mooring) where the VLCC unloads the crude oil is 75%. This study has improved the accuracy using machine learning, and the predictive accuracy of XGBoost and Random Forest with oversampling maintain 97% from one hour to 72 hours (3days). And the error rate also decreases. The errors are that predicts “close” (the VLCC could not unload crude oil because of weather) but actual case is “open” (the VLCC could unload crude oil) and predicts “open” (the VLCC could unload crude oil) but the actual case is “close” (the VLCC could not unload crude oil because of weather). The first case could cause the profit loss. In case that the oil reserve is expected to be shortage, the company could lease crude oil from KNOC(Korea National Oil Company), but the rent for crude oil is incur. If predict “close”, so lease crude oil from KNOC, but the actual case is “open”. The profit loss incur in this case. The second case could cause much more severe profit loss. If predict “open”, but the actual case is “close”. In this case, the process throughput is reduced, and it is tremendous profit loss, because the complex hydroskimming margin that price difference between product and crude oil is plus. The fall out rate ($\frac{FP}{N}$) is related to the first case. The fall out rate of existing empirical way is 60%, and the fall out rate of random forest and XGBoost is lower than existing empirical way. The miss rate ($\frac{FN}{P}$) is related to second case. The miss rate of existing empirical way is 13%, and the miss rate of random forest and XGBoost is lower than existing empirical way. Until 72 hours, the miss rate of random forest and XGBoosts is 2% and 1%, and the fall out rate of random forest and XGBoosts is 4% and 5%. They are 11~12% and 55~56% lower than existing way’s.

This study has proposed a numerical and data analytic approach to predict if the VLCC (Very Large Crude Carriers) unloads crude oil to onshore crude storage tank based on weather forecast. This proposed study is practical and used for real business. Before this study, the process of making decision if the VLCC unloads crude oil to onshore crude storage tank is based on empirical and dependent on personal decision. This study provides prediction model using meteorological data at specific area. In order to this, machine learning such as logistic regression, random forest, XGBoost and SVM is used to provide objective criteria and increase predictive accuracy. This study provides more accurate and objective model and the model can be used for the criterion of prediction if the VLCC could work based on meteorological data.

Even this study provides prediction from one hour to 3 days, the study predicting farther should be complemented by future research. The study will be useful for industry importing or exporting products by ships.

References

[1] Scher, Sebastian, and Messori, Gabriele. "Predicting Weather Forecast Uncertainty with Machine Learning." Quarterly Journal of the Royal Meteorological Society 144, no. 717 (2018): 2830-841.

[2] Lily Ingsrisawang, Supawadee Ingsriswang, Saisuda Somchit, Prasert Aungsuratana, and Warawut Khantiyanan. "Machine Learning Techniques For Short-Term Rain Forecasting System In The Northeastern Part Of Thailand." 2008.

[3] Joshi, J., C. Tankeshwar, and K. Srivastava. "Hidden Markov Model for Quantitative Prediction of Snowfall and Analysis of Hazardous Snowfall Events over Indian Himalaya." Journal of Earth System Science 126, no. 3 (2017): 1-12.

[4] Sharma, N, Sharma, P, Irwin, D, and Shenoy, P. "Predicting Solar Generation from Weather Forecasts Using Machine Learning." 2011 IEEE International Conference on Smart Grid Communications (SmartGridComm), 2011, 528-33.

[5] Dehalwar, Vasudev, et al. "Electricity load forecasting for Urban area using weather forecast information." 2016 IEEE International Conference on Power and Renewable Energy (ICPRE). IEEE, 2016.

[6] Rasouli, Kabir, William W Hsieh, and Alex J Cannon. "Daily Streamflow Forecasting by Machine Learning Methods with Weather and Climate Inputs." Journal of Hydrology 414-415 (2012): 284-93.

[7] Hamisu, Aminu A, Stephen Kabantiok, and Meihong Wang. "Refinery Scheduling of Crude Oil Unloading with Tank Inventory Management." Computers and Chemical Engineering 55 (2013): 134-47.

[8] Zhang, Jian, Yanqin Wen, and Qiang Xu. "Simultaneous Optimization of Crude Oil Blending and Purchase Planning with Delivery Uncertainty Consideration." Industrial and Engineering Chemistry Research 51, no. 25 (2012): 8453-464.

[9] Hamzah, B.A, V.L Forbes, Jalila A Jalil, and M.N Basiron. "The Maritime Boundaries of Malaysia and Indonesia in the Malacca Strait: An Appraisal." Australian Journal of Maritime & Ocean Affairs 6, no. 4 (2014): 207-26.

[10] Bouejla, Amal, Xavier Chaze, Franck Guarnieri, and Aldo Napoli. "A Bayesian Network to Manage Risks of Maritime Piracy against Offshore Oil Fields." Safety Science 68, no. C (2014): 222-30.

[11] Cheng, Lifei, and Marco A Duran. "Logistics for World-wide Crude Oil Transportation Using Discrete Event Simulation and Optimal Control." Computers and Chemical Engineering 28, no. 6-7 (2004): 897-911.

[12] Gupta, Sourabh, and Nan Zhang. Industrial and Engineering Chemistry Research 49, no. 3 (2010): 1325-332.

[13] James, Scott C, Yushan Zhang, and Fearghal O'Donncha. "A Machine Learning Framework to Forecast Wave Conditions." Coastal Engineering 137 (2018): 1-10.

[14] Solomatine, Dimitri P., and Avi Ostfeld. "Data-driven modelling: some past experiences and new approaches." Journal of hydroinformatics 10.1 (2008): 3-22.

[15] Zijderveld, Annette. "Neural network design strategies and modelling in hydroinformatics." (2003).

[16] Corzo, Gerald, and Dimitri Solomatine. "Baseflow Separation Techniques for Modular Artificial Neural Network Modelling in Flow Forecasting." Hydrological Sciences Journal 52, no. 3 (2007): 491-507.

[17] Zamani, A., Solomatine, Azimian, and Heemink. "Learning from Data for Wind–wave Forecasting." Ocean Engineering 35, no. 10 (2008): 953-62.

[18] James, Scott C, Yushan Zhang, and Fearghal O'Donncha. "A Machine Learning Framework to Forecast Wave Conditions." Coastal Engineering 137 (2018): 1-10.

[19] Allard, R., Rogers, Carroll, Rushing, and Naval Research Lab Stennis Space Center MS Oceanography Div. Validation Test Report for the Simulating Waves Nearshore Model (SWAN): Cycle III, Version 40.11, 2004.

[20] Mahjoobi, J., and A. Etemad-Shahidi. "An Alternative Approach for the Prediction of Significant Wave Heights Based on Classification and Regression Trees." Applied Ocean Research 30, no. 3 (2008): 172-77.

[21] Mahjoobi, J., and Ehsan Adeli Mosabbeb. "Prediction of Significant Wave Height Using Regressive Support Vector Machines." Ocean Engineering 36, no. 5 (2009): 339-47.

[22] Ledolter, Johannes. "Logistic Regression." In Data Mining and Business Analytics with R, 83-107. Hoboken, NJ, USA: John Wiley & Sons, 2013.

[23] Cutler, Adele, D Richard Stevens, John R Zhang, Cha Ma, Zhang, Cha, and Ma, Yunqian. "Random Forests." In Ensemble Machine Learning: Methods and Applications, 157-75. 2012 ed. Boston, MA: Springer US, 2012.

[24] Breiman, Leo. "Bagging predictors." Machine learning 24.2 (1996): 123-140.

[25] Sukhpreet Singh Dhaliwal, Abdullah-Al Nahid, and Robert Abbas. "Effective Intrusion Detection System Using XGBoost." Information 9, no. 7 (2018): 149.

[26] Chen, Tianqi, and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System." 13-17 (2016): 785-794.

[27] Fawcett, Tom. "An Introduction to ROC Analysis." Pattern Recognition Letters 27, no. 8 (2006): 861-74.

[28] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273-297.